SUPERVISED LEARNING CLASSIFICATION

TOTAL MARKS:70

DATA DESCRIPTION: This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.Age of the patient: Any patient whose age exceeded 89 is listed as being of age "90

ATTRIBUTES:

- Attribute Information: (classes: edible=e, poisonous=p)
- cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
- cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
- cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
- bruises: bruises=t, no=f
- odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
- gill-attachment: attached=a, descending=d, free=f, notched=n
- gill-spacing: close=c, crowded=w, distant=d
- gill-size: broad=b, narrow=n
- gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
- stalk-shape: enlarging=e, tapering=t
- stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
- stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
- stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- veil-type: partial=p, universal=u
- veil-color: brown=n, orange=o, white=w, yellow=y
- ring-number: none=n, one=o, two=t
- ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
- spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o,purple=u, white=w, yellow=y
- population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
- habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

1. **Read the dataset (tab, csv, xls, txt, inbuilt dataset)**
2. **Summarize important observations from the data set (5 MARKS)**

   Some pointers which would help you, but don't be limited by these

   a. *Find out number of rows; no. & types of variables (continuous, categorical etc.)*
   b. *Calculate five-point summary for numerical variables*
   c. *Summarize observations for categorical variables – no. of categories, % observations in each category*

3. **Check for defects in the data. Perform necessary actions to 'fix' these defects (5 MARKS)**

   Some pointers which would help you, but don't be limited by these

   a. *Do variables have missing/null values?*
   b. *Do variables have outliers?*
   c. *Is the Target distributed evenly? Is it a defect? If Yes, what steps are being taken to rectify the problem.*

4. **Summarize relationships among variables (10 marks)**

   a. *Plot relevant categorical plots. Find out which are the variables most correlated or appear to be in causation with Target? Do you want to exclude some variables from the model based on this analysis? What other actions will you take?*
   b. *Plot all independent variables with the target & find out the relationship? Perform the Relevant Tests to find out if the Independent variables are associated with the Target Variable.*

      Hint: based on your observations you may want to transform features or create additional features.

5. **Split dataset into train and test (70:30) (5 MARKS)**

   a. *Are both train and test representative of the overall data? How would you ascertain this statistically?*

6. **Fit a base model and explain the reason of selecting that model. Please write your key observations. (15 MARKS)**

   a. *What is the overall Accuracy? Please comment on whether it is good or not.*

b. *What is Precision, Recall and F1 Score and what will be the optimization objective keeping in mind the problem statement.*

c. *Which variables are significant?*

d. *What is Cohen's Kappa Value and what inference do you make from the model*

e. *Which other key model output parameters do you want to look at?*

7. How do you improve the accuracy of the model? Write clearly the changes that you will make before re-fitting the model. Fit the final model. (20 MARKS)

Please feel free to have any number of iterations to get to the final answer. Marks are awarded based on the quality of final model you are able to achieve.

8. Summarize as follows (10 MARKS)

1. *Summarize the overall fit of the model and list down the measures to prove that it is a good model*
2. *Write down a business interpretation/explanation of the model – which variables are affecting the target the most and explain the relationship. Feel free to use charts or graphs to explain.*
3. *What changes from the base model had the most effect on model performance?*
4. *What are the key risks to your results and interpretation?*